Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input

PNAS
www.pnas.org

Supplementary Information for

Early phonetic learning without phonetic categories
Insights from large-scale simulations on realistic input

Thomas Schatz, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao and Emmanuel Dupoux

Thomas Schatz.
E-mail: thomas.schatz.1986@gmail.com

This PDF file includes:

Supporting Information Text

Supplementary Materials and Methods.

1. Datasets. The BUC and GPJ corpora annotations present a number of inconsistencies and were curated in-house. In particular, readers for the GPJ corpus often need several takes before they read an utterance correctly and the failed takes are included in the original corpus. We only keep the final take for each sentence. For the two spontaneous speech corpora, we keep disfluencies typical of spontaneous speech (such as hesitations, word fragments, pronunciation errors, fillers, etc.), but remove parts that were not phonetically transcribed or that include other kinds of noise or silence (96.11% and 80.38% of all utterances are kept for the BUC and CSJ corpora, respectively).

Phonetic transcriptions for the two read speech corpora are obtained by combining the read text with a phonetic dictionary. For the two spontaneous speech corpora, a manual phonetic transcription of the recordings is used. Word units, which are not directly apparent in the Japanese writing system, are obtained from the phonetic transcriptions by a Japanese morphological parser for the read Japanese corpus. For the spontaneous Japanese corpus, we use the provided 'Long Word Units' as words. We exclude phonemes occurring with frequency less than 1 in 10,000 by removing any utterance in which they occur and we harmonize the transcriptions in order to have the same phonemic inventory for the read and spontaneous corpora for each language. No phonemes are excluded for the American English corpora. For the Japanese corpora, a few geminate consonants are excluded (/bː/, /zː/, /hː/, /dː/, /ʑː/, /gː/, /ɸː/ for both corpora and /tsː/ for the GPJ corpus only). The retained phonemic inventory for American English consists of 24 consonants (/p/, /t/, /k/, /b/, /d/, /g/, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /ʧ/, /ʤ/, /m/, /n/, /ŋ/, /h/, /ɹ/, /l/ /w/, /j/) and 15 vowels (/ɪ/, /iː/, /ɛ/, /ʌ/, /ɝ/, /ae/, /ɑː/, /ɔː/, /ʊ/, /uː/, /eɪ/, /aɪ/, /aʊ/, /ɔɪ/, /oʊ/). The retained phonemic inventory for Japanese consists of 27 consonants (/p/, /t/, /k/, /pː/, /tː/, /kː/, /b/, /d/, /g/, /s/, /ɕ/, /sː/, /ɕː/, /z/, /ʑ/, /ts/, /tsː/, /tɕ/, /tɕː/, /m/, /n/, /ɴ/, /h/, /ɸ/, /r/, /w/, /j/) and 10 vowels (/ä/, /e/, /i/, /o/, /ɯ/, /äː/, /eː/, /iː/, /oː/, /ɯː/). For each corpus, timestamps are obtained for the phonetic transcriptions through forced alignment with an automatic speech recognition (ASR) system (same architecture for the acoustic model as for the phoneme recognizer baseline described in Section 2 below, trained on the full corpus).

2. Phoneme recognizer baselines. As a baseline, we also train a phoneme recognizer on the training set of each corpus, with explicit supervision (i.e. providing the phonemic transcriptions of the training stimuli along with the waveforms). Specifically, we use the Kaldi toolkit (1) for automatic speech recognition (ASR) to train a hidden Markov model Gaussian mixture model (HMM-GMM) acoustic model and a phoneme-level bigram language model for each training set. The same training recipe (adapted from the Wall Street Journal corpus recipe), with the same parameters is used to train a separate model on each of the four corpora. The acoustic model takes the form of a probabilistic generative model with each phoneme modeled as a set of contextual variants that are allowed to depend on word-position and preceding and following phonemes. Each variant is itself modeled as a tri-state hidden Markov model with diagonal covariance Gaussian mixture emission probabilities. The models are adapted to speakers both during training and test through feature-space maximum likelihood linear regression (fMLLR). See the Kaldi toolkit documentation for more detail (http://kaldi-asr.org/doc/).

The trained acoustic and language models are combined (with kaldi acoustic scale parameter set to 0.1) to obtain representations of test stimuli (possibly in a 'foreign' language) under the form of a sequence of frame-level Viterbi-smoothed posterior probability vectors. We extract frame-level posterior probabilities at two granularity levels: actual phonemes—to which we refer as the *phoneme recognizer* baseline—and individual states of the contextual hidden Markov models—to which we refer as the *ASR phone state* baseline.

3. Analysis of learned representations.

Correction for possible misalignment in the acoustic (in)variance test. We compensate for possible misalignment of the central phones' central frames by allowing the dominant unit at the central frame to be replaced by any unit that was dominant at some point within the previous or following 46ms, provided this brings down the overall count of distinct dominant units for the ten occurrences. Finding the optimal way to assign dominant units under this constraint corresponds to solving an instance of the NP-complete *minimal hitting set size problem* (2). We are able to solve the problem exactly in most cases, due to the small size of the considered instances. In the few cases where we are not able to solve the problem exactly, our solver provides a lower bound on the number of representations and we use a greedy search to obtain an upper bound. Although the effect on the results is very small, we report lower bounds for the Gaussian mixture models and upper bounds for the *phoneme recognizer* and *ASR phone state* baselines, in order to be maximally conservative.

Stimulus selection for the acoustic (in)variance test. To avoid potentially mispronounced short function words and possible co-articulation effect across word boundaries, for the acoustic (in)variance test, we select only words of at least five phonemes and study their central phoneme(s).* We sample uniformly at random a subset of ten occurrences (by a single speaker or by at least ten distinct speakers, depending on the condition) for each such word with enough repetitions in the test set. We report results averaged over ten independent runs of this stimulus sampling procedure. The results are also averaged over the two possible 'central phone' positions for words of even length and—in the within-speaker condition—over all available speakers for a given word type. This yields one average number of distinct dominant units per tested word type. The number

*This stimulus selection procedure was only applied for the acoustic (in)variance test and has the effect of making the test more conservative—i.e. the learned representations would look even more variable without this restriction. Other analyses were not restricted to such words, and all model training was carried out with unfiltered continuous speech that contained words of all different lengths in unsegmented whole sentences.

of available word types matching the specified conditions is 13 (within speaker) and 476 (across speaker) for the American English test stimuli and 83 (within speaker) and 408 (across speaker) for the Japanese test stimuli. As an example, here are the word types selected for the within-speaker American English condition: unquote, billion, dollars, hundred, company, market, million, mister, nineteen, percent, seven, seventy, thousand. For the within-speaker condition, we additionally listened to each test stimulus to identify potential mispronounced, noisy or misaligned stimuli and we checked that excluding these stimuli from the analysis (0/83 word types, 4/1048 word tokens excluded for American English; 14/168 word types, 204/2217 word tokens excluded for Japanese) did not affect the overall pattern of results (Figure S8).

4. Deriving systematic model predictions. We systematically seek phonetic contrasts of American English and of Japanese for which the learning mechanism under study robustly predicts a significant cross-linguistic difference in discrimination between Japanese- and American English-learning infants. By *robust* we mean that (a) a significant difference in discrimination errors between models trained on American English and Japanese is consistently found across possible choices for the training and test registers, and (b) that the magnitude of this difference does not decrease when the amount of training input is increased. The former criterion allows us to rule out effects that would reflect peculiarities of the training and/or test stimuli rather than an intrinsic property of the language pair under study. The latter criterion allows us to rule out transient effects that might reflect peculiarities of the model initialization and/or be unlikely to be observed empirically.

We define the predicted cross-linguistic effect for a phonetic contrast as the expected difference in average ABX discrimination error between an 'American English-native' and a 'Japanese-native' model on that contrast, where the expectation is taken over the choice of American English model, Japanese model, test speaker, phonetic context, and choice of the a, b, and x acoustic tokens given the contrast, speaker and phonetic context. For each contrast, we perform statistical significance tests separately for each of the 8 possible combinations of training register for the American English model, training register for the Japanese model, and test register. We use the models trained on the $1/10^{th}$ training sets of each corpus for these significance tests, which allows us to take into account variance due to the model training procedure (including the choice of input data) in addition to that due to the choice of test stimuli. We estimate the predicted cross-linguistic effect and its variance and use those estimates to conduct asymptotic bilateral z-tests of the hypothesis that the cross-linguistic effect is different from 0. We also estimate the effects (but not the variances) using the full training sets, which allows us to test whether the observed effects increase (in absolute value) with the amount of input data. We report a robust predicted cross-linguistic effect for a contrast if each of the estimated effects for that contrast (for each of the 8 possible combination of training and test registers) is in the same direction and significantly different from 0 in our asymptotic bilateral z-test, with Benjamini-Yekutieli (3) correction for multiple correlated comparisons at level $\alpha = 0.05$; and if the estimated effect for models trained on the full training sets are in the same direction and larger in absolute value than the corresponding effects estimated for models trained on the $1/10^{th}$ subsets.

In what follows, we first formally define the predicted cross-linguistic effect for a phonetic contrast P_1, P_2. We then discuss how to estimate the effect in practice from finite samples of models trained on Japanese and trained on American English, and finite samples of test acoustic tokens from phonetic categories P_1 and P_2. Finally, we explain in detail how the statistical significance of the estimated effects can be assessed.

Effect of interest. We are interested in the predicted cross-linguistic effect for a phonetic contrast P_1, P_2, i.e. the expected difference in average ABX discrimination error between a model trained on language L_1 and a model trained on language L_2, which we denote as $\delta(P_1, P_2, L_1, L_2)$ and define formally below.[†] Let us consider a model M trained on input language L, input register R_I and input amount A_I, and tested on phonetic category P from test language L_T in phonetic context C (preceding and following phonetic category) from test speaker S with test register R_T. Let us note

$$p_{P,L,R_I,A_I,L_T,R_T}(\mathfrak{R} \mid M, S, C),$$

the probability distribution over model representations $\mathfrak{R}$, where we treat the trained model M, test speaker S and test context C as conditioning random variables and assume fixed values for the other parameters. Then, the predicted cross-linguistic effect for phonetic contrast P_1, P_2 and training languages L_1, L_2 is defined as

$$\delta(P_1, P_2, L_1, L_2) := \mathbf{E}_{M_1,M_2,S,C}[\epsilon(P_1, P_2, M_1, S, C) - \epsilon(P_1, P_2, M_2, S, C)],$$

where

- M_x for x in $\{1, 2\}$ is a randomly sampled trained model for input language L_x, training register $R_{I,x}$ and input amount $A_{I,x}$;
- S is a randomly chosen test speaker and C is a context chosen uniformly at random among available test phonetic contexts, for test language L_T, test register R_T and test phonetic contrast(P_1, P_2);
- $\epsilon(P_1, P_2, M_x, S, C)$ is the symmetric ABX discrimination error, defined as

$$\epsilon(P_1, P_2, M_x, S, C) := \frac{1}{2}[e(P_1, P_2, M_x, S, C) + e(P_2, P_1, M_x, S, C)],$$

[†] This is for a given choice of input registers $R_{I,1}$ and $R_{I,2}$ and input amounts $A_{I,1}$ and $A_{I,2}$ for each model, and of test language L_T and test register R_T (which we constrain to be the same for the two tested phonetic categories in our experiments). To avoid clutter, we do not indicate these dependencies explicitly in the notation.

with

$$e(P_1, P_2, M_x, S, C) := p[d(A, X) < d(B, X)] + \frac{1}{2}p[d(A, X) = d(B, X)],$$

for A, X drawn independently from $p_{P_1,L}(\mathfrak{R} \mid M_x, S, C)$ and B drawn from $p_{P_2,L}(\mathfrak{R} \mid M_x, S, C)$.

This is the quantity we seek to estimate, given our trained models in English and Japanese, and the particular acoustic tokens in our corpora from the phonetic categories we would like to test.

Estimation of the effect. In order to obtain a sample of model representations $S_{P,M,L_T,R_T,S,C}$ for each relevant combination of the index variables, we extract a representation of each test acoustic token for each model M.[‡] For each combination of test language L_T, test register R_T, test speaker S and test phonetic context C, we obtain a sample of up to 5 acoustic realizations of each phonetic category from the test corpus. For each combination of training language L, training register R_I, we obtain one model trained on the full training set and 10 models that are each trained on $1/10^{th}$ of it.

Given these samples from the distributions of model representations of test stimuli, we define the following estimator of $\delta(P_1, P_2, L_1, L_2)$,

$$\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2) := \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}}\frac{1}{|\mathcal{C}(S)|}\sum_{C\in\mathcal{C}(S)}\left(\frac{1}{|\mathcal{M}_1|}\sum_{M_1\in\mathcal{M}_1}\hat{\epsilon}(S_{P_1,M_1,S,C}, S_{P_2,M_1,S,C}) - \frac{1}{|\mathcal{M}_2|}\sum_{M_2\in\mathcal{M}_2}\hat{\epsilon}(S_{P_1,M_2,S,C}, S_{P_2,M_2,S,C})\right),$$

where $\mathcal{S}$ is the set of sampled test speakers, $\mathcal{C}(S)$ is the set of contexts available for the target contrast from test speaker S, $\mathcal{M}_1$ and $\mathcal{M}_2$ are the sampled models for training language L_1 and L_2 respectively and $\hat{\epsilon}$ is the estimator for the ABX discrimination error defined in the Material and Methods section of the main text.

Provided there is no systematic bias in how phonetic contexts are missing from the sample of any particular test speaker, $\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)$ can be shown to be an unbiased estimator of $\delta(P_1, P_2, L_1, L_2)$.

Significance testing. We want to assess the contrasts for which a significant cross-linguistic difference in discriminability is observed. In order to do assess significance, we need a test statistic with a known distribution. For given P_1, P_2, L_1, L_2, we define

$$\hat{D}(S, M_1, M_2) := \frac{1}{|\mathcal{C}(S)|}\sum_{C\in\mathcal{C}(S)}[\hat{\epsilon}(S_{P_1,M_1,S,C}, S_{P_2,M_1,S,C}) - \hat{\epsilon}(S_{P_1,M_2,S,C}, S_{P_2,M_2,S,C}).]$$

It is straightforward to check that

$$\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2) = \frac{1}{|\mathcal{S}||\mathcal{M}_1||\mathcal{M}_2|}\sum_{\substack{S\in\mathcal{S}\\ M_1\in\mathcal{M}_1\\ M_2\in\mathcal{M}_2}}\hat{D}(S, M_1, M_2).$$

$\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)$ can thus be interpreted as a (generalized) U-statistic with kernel $\hat{D}$ of order 3 and degree $(1, 1, 1)$ (4), applied to mutually independent i.i.d. samples $\mathcal{S}$, $\mathcal{M}_1$ and $\mathcal{M}_2$ (where an element S of $\mathcal{S}$ is effectively a sample of up to five acoustic tokens for each phonetic context available from speaker S for the target phonetic contrast).

Assuming this U-statistic is not degenerate, we can apply the central limit theorem for U-statistics (4) to obtain that

$$\frac{\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)}{\mathbf{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]}$$

has an asymptotic normal distribution with mean $\delta(P_1, P_2, L_1, L_2)$ and variance 1. Provided we can estimate the variance of the estimator $\mathbf{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$, this result allows us to perform asymptotic z-tests of $\mathcal{H}_0 :\ \delta(P_1, P_2, L_1, L_2) = 0$ versus $\mathcal{H}_1 :\ \delta(P_1, P_2, L_1, L_2) \neq 0$. We provide the required estimator $\hat{V}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)$ of $\mathbf{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$ in the next section.

Estimation of the variance of $\hat{\delta}$. The previous section showed that given an estimate $\hat{V}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)$ of the variance $\mathbf{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$, we can compute statistical significance of the estimated differences in discrimination error between languages. In this section we derive such an estimator.

We first find an expression for $\mathbf{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)]$, then derive an estimator from it. We use n_1 to denote the number of test speakers, $|\mathcal{S}|$, n_2 to denote the number of models trained on language L_1, $|\mathcal{M}_1|$, and n_3 to denote the number of models trained on language L_2, $|\mathcal{M}_2|$. We can express the variance using the standard decomposition for the variance of a U statistic (4),

$$\begin{aligned}\mathbf{Var}[\hat{\delta}(\mathcal{S}, \mathcal{M}_1, \mathcal{M}_2)] = \frac{1}{n_1 n_2 n_3}[&(n_1-1)(n_2-1)\sigma^2_{001} + (n_1-1)(n_3-1)\sigma^2_{010} + (n_2-1)(n_3-1)\sigma^2_{100}\\ &+ (n_1-1)\sigma^2_{011} + (n_2-1)\sigma^2_{101} + (n_3-1)\sigma^2_{110}\\ &+ \sigma^2_{111}]\end{aligned}$$

where σ^2_{xyz} denotes the covariance between $\hat{D}(s_1, a_1, j_1)$ and $\hat{D}(s_2, a_2, j_2)$ for two triplets (s_1, a_1, j_1), (s_2, a_2, j_2) formed of a randomly sampled combination of a test speaker, an American English model, and a Japanese model, with the subscripts x, y,

[‡] Possibly with some missing data, as not all possible phonetic contexts occur for each speaker and each phonetic category in any given test set.

and z indicating whether the two test speakers, American English models and Japanese models, respectively, are constrained to be identical (subscript 0) or not (subscript 1). For example,

$$\sigma^2_{000} = \mathbb{E}_{s_1,s_2,a_1,a_2,j_1,j_2}[\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2)] - (\mathbb{E}_{s,a,j}[\hat{D}(s,a,j]])^2 = 0;$$
$$\sigma^2_{111} = \mathbb{E}_{s,a,j}[\hat{D}(s,a,j)^2] - (\mathbb{E}_{s,a,j}[\hat{D}(s,a,j])^2;$$
$$\sigma^2_{001} = \mathbb{E}_{s_1,s_2,a_1,a_2,j}[\hat{D}(s_1,a_1,j)\hat{D}(s_2,a_2,j)] - (\mathbb{E}_{s,a,j}[\hat{D}(s,a,j])^2.$$

We now use the above variance decomposition to derive an estimator. Let us define the order 3, degree (2, 2, 2) kernel $\psi_{k_1k_2k_3}$ for some strictly positive integers k_1, k_2, k_3, as follows

$$\begin{aligned}\psi_{k_1k_2k_3}(s_1,s_2,a_1,a_2,j_1,j_2) := \frac{1}{k_1k_2k_3}[&(k_1-1)(k_2-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))\\ &+(k_1-1)(k_3-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_1,j_2) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))\\ &+(k_2-1)(k_3-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_2,j_2) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))\\ &+(k_1-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_1,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))\\ &+(k_2-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_2,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))\\ &+(k_3-1)(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_1,j_2) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))\\ &+(\hat{D}(s_1,a_1,j_1)\hat{D}(s_1,a_1,j_1) - \hat{D}(s_1,a_1,j_1)\hat{D}(s_2,a_2,j_2))]\end{aligned}$$

Let us consider some arbitrary orderings $(s_1,s_2,...,s_{n_1})$, $(a_1,a_2,...,a_{n_2})$ and $(j_1,j_2,...,j_{n_3})$ of $\mathcal{S}$, $\mathcal{M}_1$, and $\mathcal{M}_2$, respectively. Let us also note $(n\ k)$, for any integers n and k, the set of all integer k-tuples $(i_1,i_2,...,i_k)$ such that $1 \le i_1 < i_2 < ... < i_k \le n$.

It is straightforward to show that $\psi_{n_1n_2n_3}$ is an unbiased estimator for $\mathbf{Var}[\hat{\delta}(\mathcal{S},\mathcal{M}_1,\mathcal{M}_2)]$, leading to the following symmetric unbiased estimator based on all of the available data

$$\hat{V}(\mathcal{S},\mathcal{M}_1,\mathcal{M}_2) := \frac{1}{\binom{n_1}{2}\binom{n_2}{2}\binom{n_3}{2}} \sum_{\substack{i_1,i_2\in(n_1\ 2)\\ j_1,j_2\in(n_2\ 2)\\ k_1,k_2\in(n_3\ 2)}} \psi^S_{n_1n_2n_3}(s_{i_1},s_{i_2},a_{j_1},a_{j_2},j_{k_1},j_{k_2}),$$

where $\psi^S_{n_1n_2n_3}$ is the symmetrized version of $\psi_{n_1n_2n_3}$

$$\psi^S_{n_1n_2n_3}(s_1,s_2,a_1,a_2,j_1,j_2) := \frac{1}{(2!)^3} \sum_{\substack{i_1,i_2\in S_2\\ j_1,j_2\in S_2\\ k_1,k_2\in S_2}} \psi_{n_1n_2n_3}(s_{i_1},s_{i_2},a_{j_1},a_{j_2},j_{k_1},j_{k_2}),$$

with $S_2 = \{(1,2),(2,1)\}$ the set of all permutations of $\{1,2\}$.

With this estimator for the variance of $\hat{\delta}(\mathcal{S},\mathcal{M}_1,\mathcal{M}_2)$, we can now conduct a z-test over the test statistic defined in the previous section to compute statistical significance of cross-linguistic discrimination differences.

Supplementary Discussion.

1. Input idealization in computational modeling of early phonetic learning. Modeling studies investigating the feasibility of potential learning mechanisms for early phonetic learning have typically relied on input idealizations that sidestep the lack of invariance problem and the phonetic segmentation problem, and cannot therefore alleviate the feasibility concerns related to these problems. In initial modeling work investigating the feasibility of learning phonetic categories through distributional learning (5–9), the phonetic category segmentation problem was either simply assumed to have been solved (7–9), or the input speech was assumed to consist of exemplars from a restricted number of pre-segmented or isolated syllable types, that were furthermore chosen such that automatic segmentation of the vowel nucleus based on voicing cues would be easy (5, 6). The impact of the lack of invariance problem was minimized by artificially limiting the variability of the input. Specifically, the input speech signal was: chosen from a restricted set of phonemes (5–9); occurring in a restricted set of phonetic contexts (5–7); uttered by a (very) restricted set of speakers (5, 9); available to the learner in a manually encoded (7–9) and/or restricted (5–9) phonetic feature space; drawn from synthetic parametric sound distributions fitted to corpus data rather than using corpus data directly (7, 8). Subsequent studies considered slightly more realistic variability and found that distributional learning was not sufficient anymore to learn phonetic categories accurately (10–16) and proposed additional learning mechanisms tapping into other sources of information plausibly available to infants to complement distributional learning. However, demonstrations of feasibility for the proposed mechanisms still assumed the phonetic category segmentation problem to be solved (10–12, 14–16) and/or did not fully address the lack of invariance problem by not considering the full variability of natural speech (10–16). Specifically, input speech signal was: chosen from a restricted set of phonemes (10–12, 14–16); occurring in a restricted set of phonetic contexts (12, 14, 16); uttered by a very restricted set of speakers (10, 11, 13, 15, 16); available to the learner in a manually encoded (9, 10, 12, 14–16) and/or restricted (10–12, 14–16) phonetic feature space; drawn from synthetic parametric sound distributions fitted to corpus data rather than using corpus data directly (11–14). Existing attempts to extend some

of these results to more realistic learning conditions have failed (17, 18). The few studies that attempted to model infant phonetic learning from naturalistic, unsegmented speech input remained inconclusive for lack of a suitable evaluation method (19, 20). Finally, we know of only one demonstration of feasibility for an account of early phonetic learning in which the outcome of learning is not phonetic categories (21). It also assumes the phonetic category segmentation problem to be solved and minimizes the impact of the lack of invariance problem by artificially limiting the variability of the input speech.

Modeling assumptions are necessary in any model—for example, our approach ignores the visual component of speech and uses adult-directed rather than child-directed speech—but they should be critically examined to assess their suitability relative to the research objectives. For example, whereas the assumptions typically made in previous studies were all geared toward making the learning problem easier—by sidestepping the lack of invariance and phonetic segmentation problems—we focus, as much as possible, on modeling assumptions that make it harder. This means that in our framework, positive feasibility results constitute much stronger evidence. Our framework is not devoid of modeling assumptions that make the learning problem easier; for example, we consider speech input consisting of speech from a single speaker at a time, captured by a close-range microphone, and with no overlap with environmental sounds. However, we make many fewer such simplifying assumptions than previous models and we are careful not to sidestep the phonetic category segmentation and the lack of invariance problems in particular. This ensures that our simulations are suitable to address feasibility concerns related to these problems.

2. Model initialization, learning procedure and convergence. Following Chen et al. (22), the parameters of our Gaussian mixture models are learned through the exact Markov chain Monte-Carlo (MCMC) sampling algorithm proposed in Chang & Fisher (23). This algorithm combines, in a principled way, Gibbs sampling of the parameters of instantiated mixture components (i.e. the clusters with non-empty membership at any given point in the algorithm execution) with sampling of split and merge moves that increase or reduce the number of instantiated mixture components. It is designed to combine good statistical convergence properties with computational efficiency, and in particular to allow the parallelization of the computations to accommodate large training datasets.

We also follow Chen et al. (22) for model initialization. They used the default initialization procedure in the implementation proposed by Chang & Fisher (23), which consists of assigning each data point in the training set uniformly at random to one of ten initial clusters. The mean vector and covariance matrix for each of these ten initial clusters is then taken as the mean and covariance of the points assigned to that cluster. The weights of each of the cluster in the initial mixture is obtained by drawing from a Dirichlet distribution with ten categories and concentration parameter whose i-th component, for $1 \leq i \leq 10$, is the number of points that were initially assigned to the i-th cluster.

In theory, the initial state should not influence the learning outcomes when using this algorithm. The sampling algorithm we use comes with the usual guarantees (for sampling algorithms) of global convergence to the true posterior in the limit (23), so that in principle, the initialization procedure should not matter if we run the sampling procedure for long enough. The main issue in practice is that there is usually no definitive way to determine when it has been 'long enough'. In our case, we look at the number of learned categories as a function of the number of sampling iterations (Figure S11). We see that this number is largely stabilized after about 600 iterations for all the models we train. This suggests that training the models for 1500 sampling iterations (per parameter), as we do—again following the example of (22)—is sufficient for model convergence. We also see that cross-linguistic differences emerge quite robustly on independent runs for models trained on one to two hours of speech input (Figure 3(b)). Thus, we are reasonably confident that the models have converged.

Still, we cannot completely rule out the possibility that running the algorithm for longer might ultimately lead to a different outcome (e.g. to units corresponding to phonetic categories), and that a different setting of the initial state might lead to that outcome faster. This leads us to consider the biological and psychological plausibility of the initialization procedure we used.

A prominent proposal in the literature (see 24, for example)—motivated by observations of a certain 'language-readiness' of the human brain at birth and even before (25)—is that infants start with an innately specified, 'universal' mapping from an auditory space to a phonetic space, which is then progressively altered as they gain experience with their native language. However, there have not yet been proposals for a concrete implementation of such a mapping (although see 26, for a possible technical solution).

This view is not universally shared. An alternative hypothesis has been argued to be fully compatible with the empirical record (e.g. 27, 28), according to which the observation of 'universal' phonetic discrimination abilities in newborns would correspond to an initial mode of perception of a purely auditory nature, in the absence of any mapping to phonetic space. Under this view, phonetic representations would be initiated through some form of random mapping, and subsequently refined through experience-dependent plasticity. One benefit of this latter view is that it assumes less in terms of what needs to be genetically specified than an innate universal mapping between acoustic and phonetic space.

As discussed in the main text, MFCC input features can be interpreted as the output of a (very) simple model of the peripheral auditory system, and our approach to initialization can thus be understood as an implementation of this latter view. We are not aware of many empirical constraints on what would constitute a plausible random initialization of the phonetic clusters within this auditory space, and our initialization procedure represents one possible, albeit admittedly arbitrary, solution.

3. Interpretation of simulated discrimination errors and relation to empirical observations. To evaluate our trained models, we expose them to appropriate test stimuli (e.g. exemplars of [ɹ] and [l]) and simulate discrimination tasks using the models' representation of these stimuli. Here, we discuss our criteria to decide if the models successfully account for early phonetic learning on the basis of the resulting discrimination errors. For the purpose of this article, we deem our models successful if they can account for the

cross-linguistic differences in discrimination abilities observed in infants in the first year of life for the Japanese/American English language pair we study.

The results to be accounted for come from a 2006 study by Kuhl and colleagues (29), since we are not aware of other studies directly comparing the phonetic discrimination abilities of Japanese and American English infants in the first year. Using a conditioned head turning paradigm, they found no significant difference between American English and Japanese infants' ability to discriminate a synthetic [ɹa] stimulus from a synthetic [la] stimulus at 6-8 months. Both groups answered correctly on about 65% of test trials. In contrast, at 10-12 months, American English infants were found to to be significantly more accurate than Japanese infants in the same task. American English infants answered correctly on about 75% of trials while Japanese infants answered correctly on about 60% of trials. All four groups discriminated the stimuli significantly above chance. When comparing across ages, American English 10-12 month olds were found to be significantly better at discriminating the stimuli than their 6-8 month old counterparts, whereas Japanese 10-12 month olds were not found to be significantly worse than their 6-8 month old counterparts (but see 30). We adopt the standard interpretation that these results reflect infants' discrimination of the [ɹ]-[l] contrast, and not just of the two specific stimuli tested in the experiment. We therefore test our models both on those specific stimuli (Figure S6), and on other instances of [ɹ] and [l] (Figure 3). However, we do not assume these observations of early phonetic learning in infants to mean that 10-12 month old infants have formed adult-like representations; while this is a common view in the literature, it is premised on the phonetic category hypothesis we are contesting. In particular, we do not take the results from Kuhl et al. (29) to necessarily indicate that Japanese 10-12 month olds have become nearly deaf to the [ɹ]-[l] distinction, or that American English 10-12 month olds learned to discriminate it perfectly.[§]

Given our current state of knowledge about infant cognition, there are some quantitative aspects of these results that we cannot hope to model, even in principle. First, we cannot hope to model the quantitative values of the error rates or d' measurements characterizing infant discrimination in these experiments, as these values depend strongly on the specifics of the experiments in ways that are not well understood (32). This uncertainty might potentially be accounted for through free parameters in the model, but fitting those parameters would not be feasible due to the limited number of datapoints available to constrain them.[¶] Second, we do not know the precise correspondence between an infant of a particular age and a model presented with a particular amount and quality of data. The quality and quantity of data in infants' environments does not directly translate into their *intake* (33), the data they use for learning. In addition, some of the differences in infants' behavior at different ages might also stem from developmental factors not directly related to perception, and these are not included in our model. Moreover, we do not know whether infants rely solely on learned representations for discrimination, even when those representations are just starting to be formed and might be unreliable, or whether they initially rely on language-universal input features for discrimination, and then smoothly transition to relying on the learned language-specific representations as the amount of training data increases. This prevents us from interpreting the change in discrimination errors as a function of the amount of training input given to the model on Figure 3(b) directly as a developmental trajectory for example.

Because we cannot hope to get a quantitative match in either the absolute discrimination scores or the absolute quantity of training data, we focus on modeling qualitative aspects of the empirical results. This means showing that American English models discriminate [ɹ] and [l] better than Japanese models do. We find this qualitative effect both with the original stimuli from Kuhl et al. (29), and with a broader set of speech stimuli drawn from American English speech corpora. Figure S6 shows that with small amounts of training data, the dissimilarity between the two original stimuli is roughly similar for all models. As the amount of training data increases, the two stimuli become more dissimilar for the American English models, while their dissimilarity stays roughly the same for the Japanese models. When tested on a broader set of [ɹ] and [l] stimuli, all models get better at discriminating this contrast as the amount of training data increases, but a clear cross-linguistic difference nevertheless emerges (Figure 3(b)). As noted above, there are a number of reasons why the direction of change in absolute error rates might not be reliable; but in both simulations, the increasing separation between English and Japanese models with increasing training data qualitatively matches the empirical pattern.

A limitation of this study is that it focuses on one language pair, limiting the relevant empirical record to mostly one study (29). Mugitani and colleagues (34) suggested that vowel length perception at 10 months could be similar in American English and Japanese listeners; our models appear broadly consistent with this hypothesis, as we find no systematic difference in Japanese vowel length discrimination between the Japanese and American English models (see Supplementary Discussion 5). However, we do not focus on this result, as Mugitani and colleagues (34) did not directly test American English 10 month olds, and recent evidence suggests that the development of vowel length perception, for Japanese listeners at least, might be more complicated than once thought (35). As argued in the main discussion, in the longer term our modeling framework will allow evaluating the proposed learning mechanism against the empirical record on further language pairs, comparing it with other possible learning mechanisms, and designing empirical tests of their predictions.

We are not aiming to model adult data, nor are we able to interpret absolute error rates relative to infant data. Thus, the absolute levels of the discrimination errors we obtain have little bearing on our main conclusions. However, it is still interesting to get a sense of how those absolute error rates might be interpreted. To this end, we added a supervised phoneme recognizer baseline as a possible approximation of an adult-like state,[‖] In general, the supervised baselines show larger cross-linguistic differences than our (unsupervised) models do. For the [ɹ]-[l] contrast, for example, the absolute difference in discrimination errors between 'native' and 'non-native' models is about four times as large for the supervised phoneme recognizers as for the

[§] This view is supported by empirical evidence that American English infants' perception of [ɹ]-[l] develops well beyond the first year of life (31).

[¶] One potential solution might be to pool infant data across many experiments to try and calibrate task models. However, it is unclear whether this strategy could be successful, because of the heterogeneity in the way infant experiments are carried out in practice.

[‖] This is different from its role in Figures 4, S7, S9 and S10, where it is used as a possible embodiment of the linguistic notion of phonetic category.

unsupervised models. These larger crosslinguistic differences are driven by decreased performance of the supervised baselines on the 'non-native' language and increased performance on the 'native' language (Figures S3, S5), though improvement on the 'native' language does not appear robust to a register change (Figure S3). These results show that the proposed learning mechanisms for early phonetic learning is compatible with the view that one-year-olds have not yet formed mature, adult-like speech representations.[**]

We additionally included an unlearned 'auditory' input features baseline (with distances computed directly between sequence of MFCC input vectors) in Figures S3, S5, as a possible approximation of discrimination on the basis of a language-universal auditory representation. This baseline performs surprisingly well relative to both the supervised baseline and the unsupervised models in discriminating some phonetic contrasts. On average, the 'native' models do better than the baseline, and the 'non-native' models do worse, as expected (Figure S3). However, this is not true for every contrast, as can be seen for [ɹ]-[l] and [w]-[j] on Figure S5. There are a number of possible ways to interpret this result.[††] This might reflect a shortcoming common to both the unsupervised models and supervised baselines for these contrasts. It might also be that, in order to catch up with the input features baseline, our models require larger amount of training input (Figure 3(b)) or input that is more similar to what infants hear (38). Finally, another possibility is that high level language-specific representation might need to be combined with information-rich auditory representation (39) to enable accurate phonetic discrimination of certain contrasts—as appears to be the case in humans (40).

4. Interpretation and plausibility of the learned representations. It might seem surprising for infants to be learning—as part of the language acquisition process—units such as those we find, with no established linguistic interpretation. Given the relative evolutionary recency of the language faculty in humans (41), however, early phonetic learning might be grounded in domain-general perceptual learning mechanisms (42, 43), the outcome of which might not conform to a purely linguistic interpretation. Supporting this view are observations of early perceptual attunement in other modalities than speech perception—for example in face (44), voice (45), pitch (46, 47), music (48) and linguistic sign (49) perception—and in other animals than humans—for example for conspecific vocalizations in rats (50), for music in mice (51) and for faces in macaques (52). Furthermore, there is evidence that the physiological mechanisms governing the onset and offset of perceptual attunement might be similar in these different modalities and conserved from mouse to man (53–55). Furthermore, from a more adaptive/functional point of view, phonetic categories embody sophisticated linguistic knowledge and inferring them from scratch might simply be too difficult. The learned representations under the proposed account support remarkably accurate discrimination of native language word-forms (22, 56–58)—a criterion for which early phonetic representations have been proposed to be optimized (59–61). They could thus serve as a more robust intermediate point in a bootstrapping process (62) ultimately leading to language proficiency.

Another question that arises is whether the learned representations are biologically and psychologically plausible given their relatively high dimensionality—between 444 and 899 learned categories, with posterior probability vectors of matching dimension. It is questionable whether infants—or even adults—would be able to explicitly access and manipulate such detailed representations of the phonetics of very short stretches of speech. We believe, however, that the learned units are plausible at least as lower-level perceptual representations. Such high-capacity intermediate representations are commonly postulated in other domains of adult and infant cognition—for example, as part of the 'core' object recognition and the 'core' spatial navigation systems (63), with correspoding computational models typically featuring representations in even higher dimensions than the ones we consider here (64–67). Computation over such high-capacity representations is likely to be costly and might be limited to a restricted set of operations—including the formation of integrated similarity or familiarity judgments, for example. Such representations are typically seen as supporting the operation of largely subconscious cognitive processes and allowing the formation of higher-level, lower-capacity, representations over which computations can be carried out more flexibly (see 68, for example).

5. Systematic model predictions. We provide a concrete demonstration of our framework's ability to link accounts of early phonetic learning to systematic predictions regarding the empirical phenomenon they seek to explain by reporting in Table S1 phonetic contrasts of Japanese and American English for which the distributional learning mechanism we study robustly predicts a significant difference in discrimination abilities between learners of those languages. Note that nothing in our method—which we present in detail in Supplementary Materials and Methods 4—is specific to the particular distributional learning mechanism studied in this article. It applies directly to any learning mechanism taking actual speech signal as input, as long as a reasonable way to measure the (dis)similarity between the learned representations of relevant test stimuli can be provided.

Reassuringly, we find that American English [ɹ]-[l] is among the contrasts robustly predicted to be significantly harder to discriminate for Japanese-learning infants. Only two other contrasts of American English are predicted to be robustly harder to discriminate for Japanese-learning infants, both involving the rhotacized vowel [ɝ]. We are not aware of empirical comparisons of Japanese- and American English-learning infants (and even adults) having been carried out so far for these contrasts. No contrast of Japanese is predicted to be robustly harder for American-English-learning infants.

6. Advantages of our approach over traditional approaches to making predictions. Our approach to linking a learning mechanism to systematic predictions regarding infant phonetic discrimination relies on explicit simulations of the learning process. Such simulations have been carried out before (5–16, 19–21, 69), however this never resulted in concrete predictions regarding

[**] This view is supported among other things by evidence of continued phonetic learning well after the first year (see e.g. 31, 36, 37).

[††] We do not attempt to decide between these possible interpretations here, as this is not directly relevant to our main conclusions.

infants' discrimination abilities. One reason is that previous simulation studies were conducted in the context of *outcome-driven* approaches and therefore focused on testing whether phonetic categories could be learned, rather than on predicting discrimination patterns observed in infants. There are also methodological limitations that would have have severely limited the possibility of obtaining systematic predictions in these studies. One of them is the drastically simplified input used in most studies. Influences of the phonetic context on cross-linguistic differences in discrimination abilities (70) might fail to be captured when the training data is restricted to just a few contexts, for example. Or meaningful predictions might be impossible for non-native contrasts falling into part of the phonetic space that is not represented in the input when it contains only a subset of the phonetic categories of the training language (e.g. if the input consists exclusively of vowels represented in terms of their formant frequencies). Even for the studies that did attempt to model infant phonetic learning from realistic speech input (19, 20), the lack of a suitable evaluation method to handle the complex speech representations typically produced by algorithms learning from raw speech without supervision would have prevented the derivation of systematic predictions. Indeed, as we already noted, traditional signal detection theory models of discrimination tasks (71) cannot handle high-dimensional input representations, while more elaborate Bayesian probabilistic models (72) typically have too many free parameters to be practical. Moreover, traditional evaluation methods for representation learning algorithms from the machine learning literature typically assess performance on downstream tasks such as supervised classification, or against known cluster labels, rather than on the discrimination abilities measured in infants. Finally, the procurement of appropriate test stimuli for all the phonetic contrasts for which predictions are to be obtained, and the need for a sound statistical methodology to separate signal from noise in the large number of resulting predictions, would have presented two additional challenges.

In principle, an alternative to our mechanism-driven approach would be to obtain predictions by relying on pre-specified notions of the outcome of learning. In phonetic category accounts, for example, predictions could be made based on how the phonetic categories from the test language map onto the phonetic categories of the native language. This has been the standard approach in the field until now, but to the best of our knowledge, has never resulted in the kind of systematic predictions we report here. Its scalability is limited by two central difficulties related to the intrinsic complexity of the speech signal. First, given that detailed aspects of the speech signal can strongly affect discrimination abilities (70, 73), making systematic predictions would require extraordinarily detailed phonetic descriptions of the whole phonetic space in all of the relevant languages. Such descriptions are not available at the required scale at present, and conducting detailed phonetic analyses to obtain them would represent a colossal undertaking. Second, even on a small scale, how to carry out the required phonetic analyses is not clear. Arbitrary decisions would have to be made, for example, regarding which phonetic dimensions to include, how to characterize these dimensions acoustically, how to characterize discrete categories in the presence of gradient effects, and how to concretely relate the observed cross-linguistic phonetic differences to predicted discrimination abilities. Some of this methodological uncertainty has been sidestepped in practice by relying on empirical assimilation patterns—adults' judgments regarding what sound from their native language is most similar to a non-native stimulus—to guide the derivation of predictions in an ad hoc fashion. This is not a scalable solution, however, given the costs associated with human experimentation. It also fails to explain how the observed assimilation patterns arise in the first place.

Our modeling framework provides the first practical, scalable way to link accounts of early phonetic learning to systematic predictions regardings infant phonetic discrimination. Key innovations underlying the success of our framework relative to previous approaches include a focus on mechanisms rather than outcomes, and on mechanisms capable of learning from naturalistic speech in particular, resulting in models capable of making systematic predictions. The testing of these models at scale relies on further important innovations. One of them is the use of large forced-aligned databases of transcribed continuous speech recordings to procure relevant test stimuli. Another is the use of the machine ABX test to link model representation of test stimuli to concrete, systematic predictions regarding infants' discrimination abilities. The machine ABX test is an automatized, parameterless measure of discriminability that is computationally tractable, statistically efficient, and can handle representations in essentially any format, as long as a reasonable way to measure the similarity between the speech representations to be evaluated can be provided, making it easy to compare the predictions from different models (74). The rationale for such an evaluation method, with a focus on simplicity of use and scalability—rather than seeking to provide a detailed model of infants' behavior in a particular experimental paradigm—is the idea that different discrimination tasks all index a common perceptual process and should result in qualitatively similar discrimination patterns—an idea that has received empirical support from the signal detection literature (71). Finally, another important innovation is the careful statistical analysis—taking into account noise sources in both model training and evaluation (see Supplementary Materials and Methods 4)—which allows us to tease out reliable effects in the large number of generated predictions.

Table S1. Phonetic contrasts for which a significant difference in discriminability between American English- and Japanese-learning infants is *robustly* predicted by the proposed distributional learning mechanism. That is, for each possible choice of training and test register, these contrasts show a significant difference in discrimination errors between models trained on American English and Japanese, and the magnitude of this difference does not decrease as the training data size is increased. See Supplementary Materials and Methods 4 for justification of these criteria and details of the method.

Language	Contrast	Easier for learners of	Average difference in discrimination error
Am. English	[ɝ] - [ɪ]	Am. English	5.4%
Am. English	[ɝ] - [ʌ]	Am. English	4.8%
Am. English	[ɹ] - [l]	Am. English	3.7%

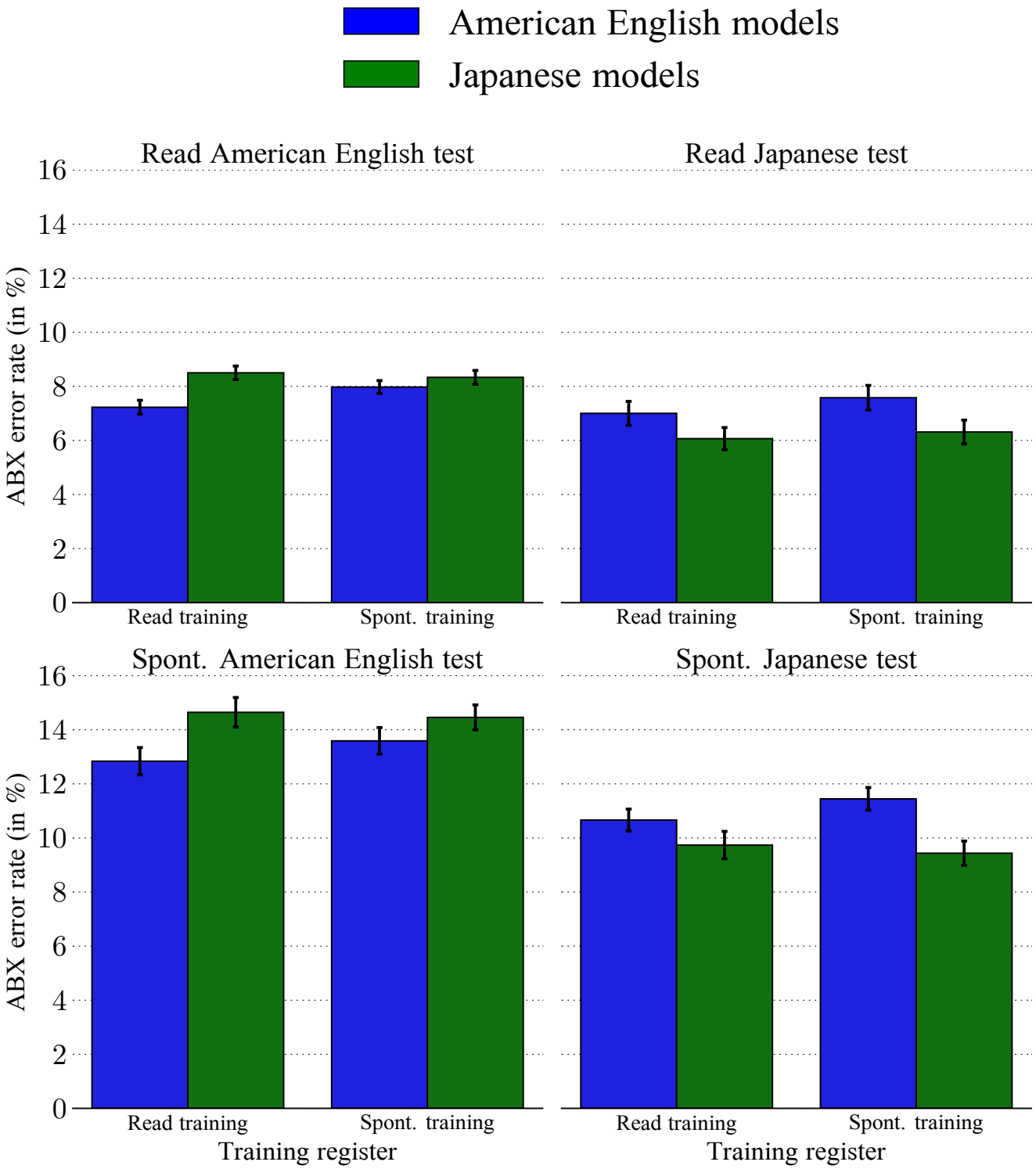

Fig. S1. Average ABX error rates over all consonant and vowel contrasts obtained with each of our four Gaussian mixture models on each of the four test sets. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. On all four test sets, 'native' models make fewer discrimination errors than 'non-native' models, illustrating the robustness of the observed native advantage.

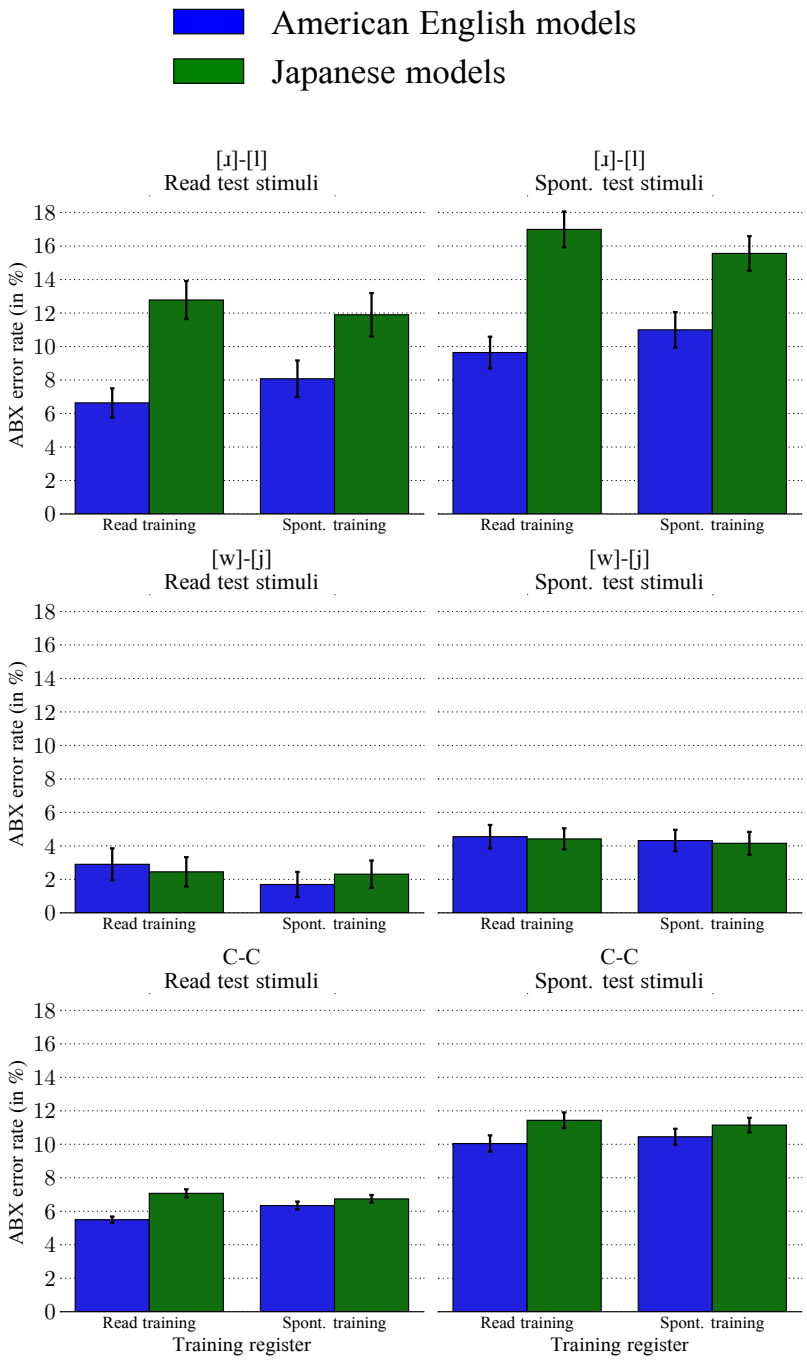

Fig. S2. ABX error rates for the American English [ɹ]-[l] contrast and two controls: American English [w]-[j] and average over all American English consonant contrasts. Error-rates are reported for each of the four trained Gaussian mixture models and each of the two American English test sets. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. Results show that the specific deficit for American English [ɹ]-[l] discrimination for 'Japanese' models compared to 'American English' models is robustly observed across all training and test conditions.

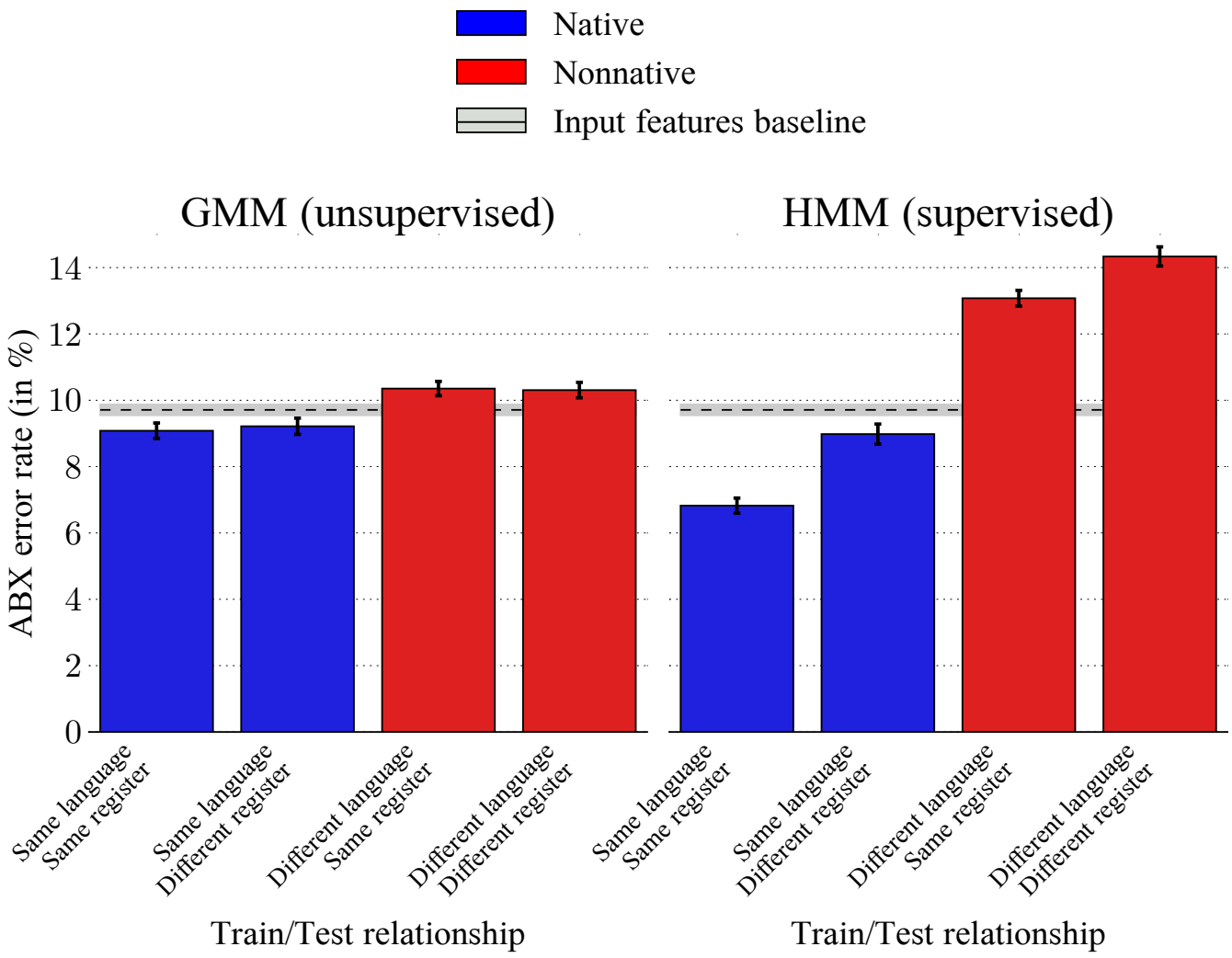

Fig. S3. Average ABX error rates over all consonant and vowel contrasts obtained with unsupervised Gaussian mixture models (GMM), with a supervised phoneme recogniser baseline (HMM) and with an input features (MFCC) baseline, as a function of the match between the training set and test set language and register. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. For both Gaussian mixture models and the phoneme recogniser baseline, the 'Native' (blue) conditions, with training and test in the same language, show fewer discrimination errors than the 'Non-native' (red) conditions. Also, in both cases the 'Native' conditions show fewer errors than the input features baseline, while 'non-native' conditions show more errors. However, the native language effect (difference between 'native' and 'non-native' models) is bigger for the supervised than the unsupervised models. Also, whereas the unsupervised models generalise very well across registers, the supervised models appear to overfit the training register.

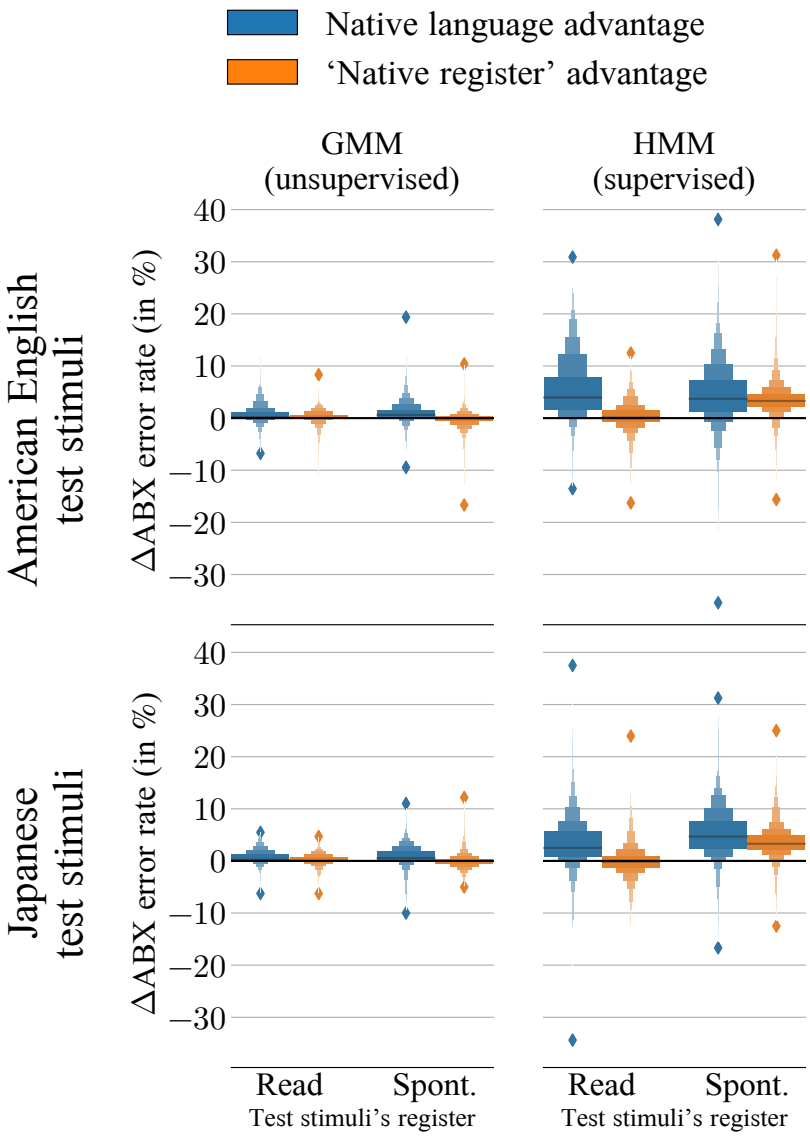

Fig. S4. Letter-value plots(75) of the distribution of 'native' advantages across all tested phonetic contrasts (pooled over both languages) for the unsupervised Gaussian mixture models (GMM) and the supervised phoneme recogniser baseline (HMM). The native language advantage is the increase in discrimination error for a contrast of language L1 between a 'L1-native' model and a model trained on the other language, keeping the training register constant. The 'native register' advantage is the increase in error for a contrast of register R1 between a 'R1-native' model and a model trained on the other register, keeping the training language constant. For both types of models and in all tested cases, the reduction in the average discrimination error between 'native language' and 'non-native language' conditions is not driven by just a few contrasts. The 'native register' only seems to play a role for the supervised models. In particular supervised models trained on read speech appear to have trouble discriminating spontaneous speech stimuli, while supervised models trained on spontaneous speech do not have problem discriminating read speech stimuli.

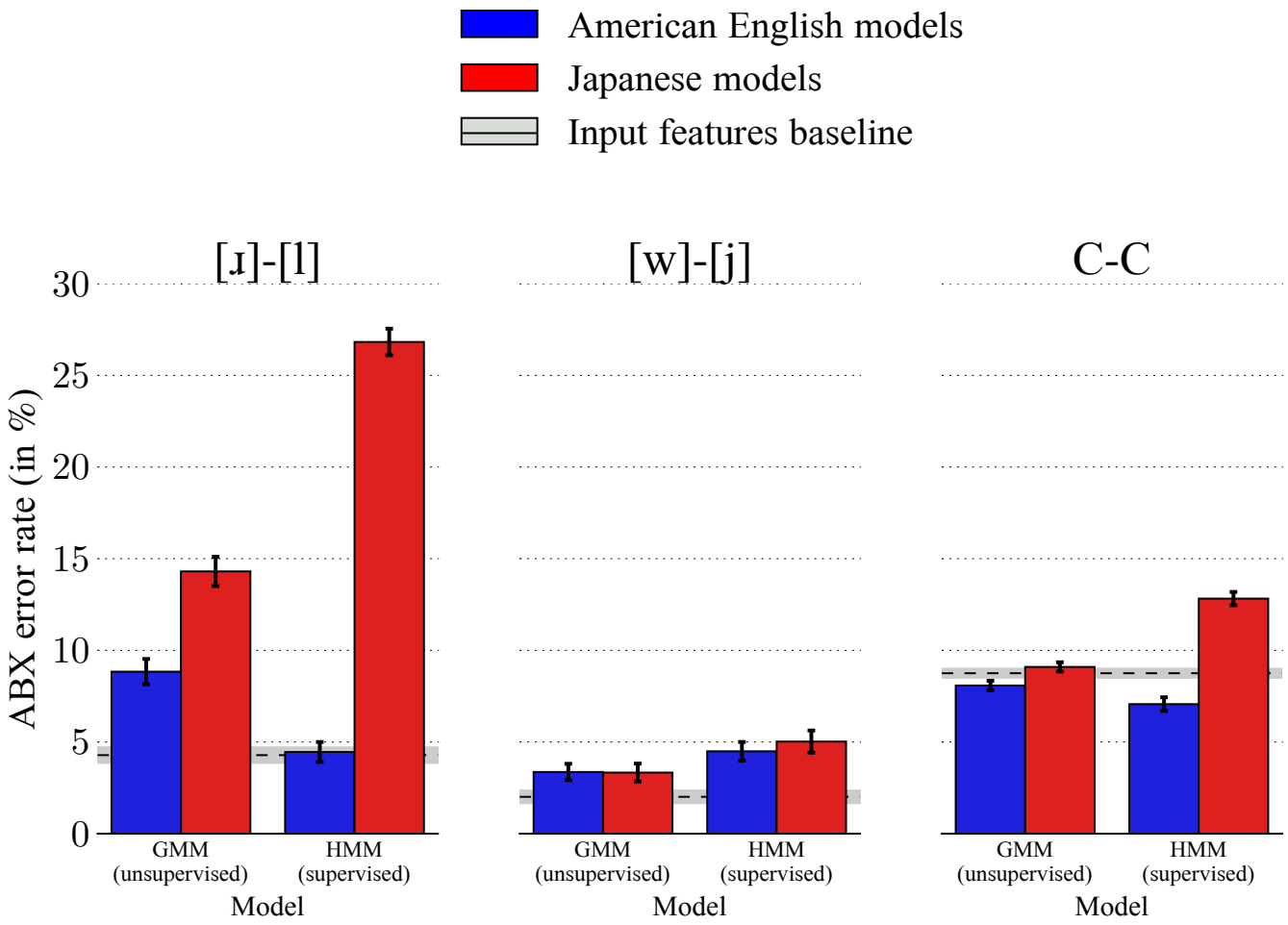

Fig. S5. ABX error rates for the American English [ɹ]-[l] contrast and two controls: American English [w]-[j] and average over all American English consonant contrasts (C-C). Error rates averaged over the two American English test sets and across model's training registers are reported for the unsupervised Gaussian mixture models (GMM), the supervised phoneme recogniser baseline (HMM) and the input features baseline. Error bars correspond to plus and minus one standard deviation of the errors across resampling of the test stimuli speakers. The specific deficit for American English [ɹ]-[l] discrimination for 'Japanese' models compared to 'American English' models is observed with both the unsupervised Gaussian mixtures and the supervised phoneme recognisers. The size of the deficit is larger for the supervised baseline, though, which we can interpret as the unsupervised GMM models producing somewhat immature representations of speech, like those of human infants (36), while the supervised HMM models produce more adult-like representations. Another interesting result is that the supervised American English models ('native' condition, in blue) do not outperform the input features baseline in the supervised case and underperform it in the unsupervised case. This suggests that some of the detailed information relevant to discrimination that was present in the input features was not preserved through the learning of a different representation of the speech signal in terms of discrete Gaussian components (see Supplementary Discussion 3 for further discussion).

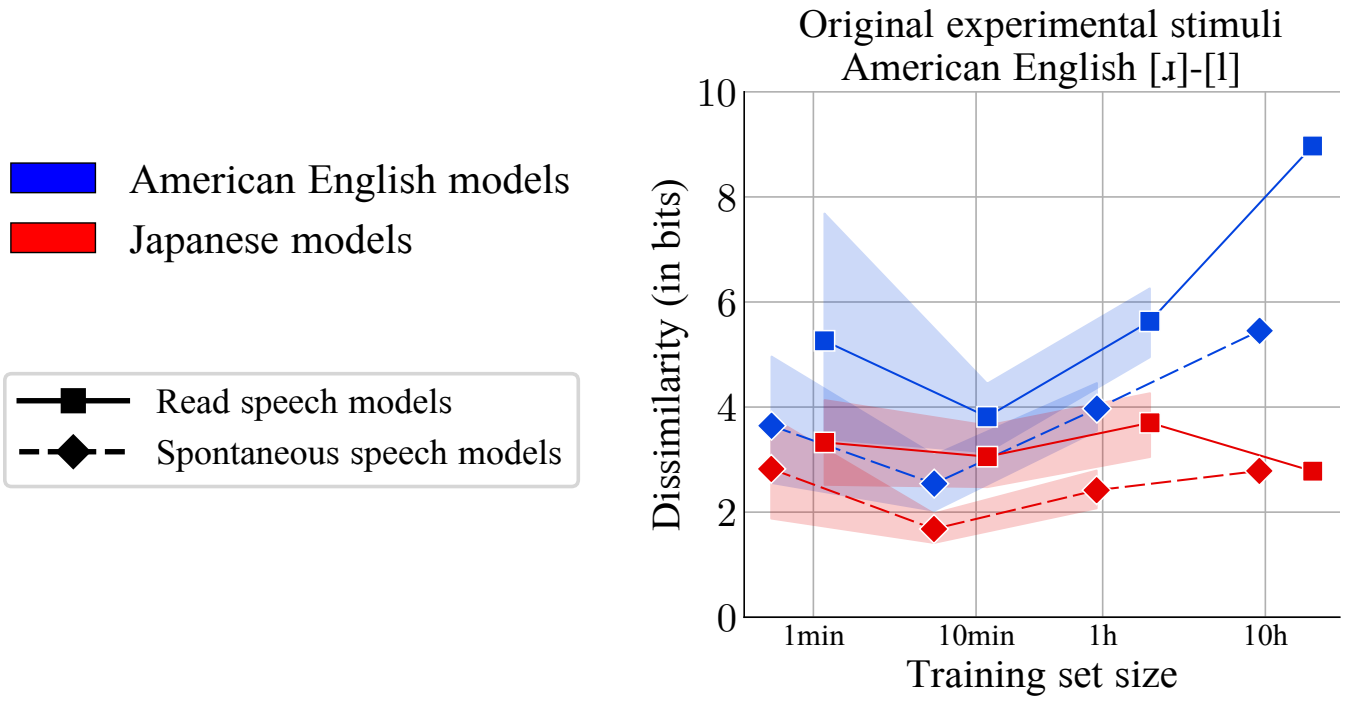

Fig. S6. Dissimilarity between the trained models' representation of a synthetized /ra/ stimulus and a synthetized /la/ stimulus as a function of the amount of input. These stimuli are those used in the empirical study which showed the emergence of a cross-linguistic difference in discriminability of these stimuli between Japanese- and American English-learning infants (29). For each selected duration (except when using the full training set), ten independent subsets are selected and ten independent models are trained. Solid lines indicate the average dissimilarity, with error bands indicating plus or minus one standard deviation. The dissimilarity corresponds to the average of the Kullback-Leibler divergence between posteriorgram representations of the stimuli along the dynamic time warping alignment path, expressed in bits (see Material and Methods). As the amount of input data increases, there does not appear to be much of a change in the dissimilarity of the two stimuli for the Japanese models, whereas there is sharp increase in dissimilarity for the American English models, especially between the 1-2h and 10-20h of training input. This is remarkably consistent with the empirically observed behavior of infants tested with these stimuli: no significant change was observed in the ability of Japanese-learning infants to discriminate these stimuli between 6-8 and 10-12 months of age, whereas American English infants became better at it (29). The predicted cross-linguistic difference between American English and Japanese learners appears to require more input to be observed reliably when testing the models with synthetic stimuli than with natural stimuli (cf. Figure 3).

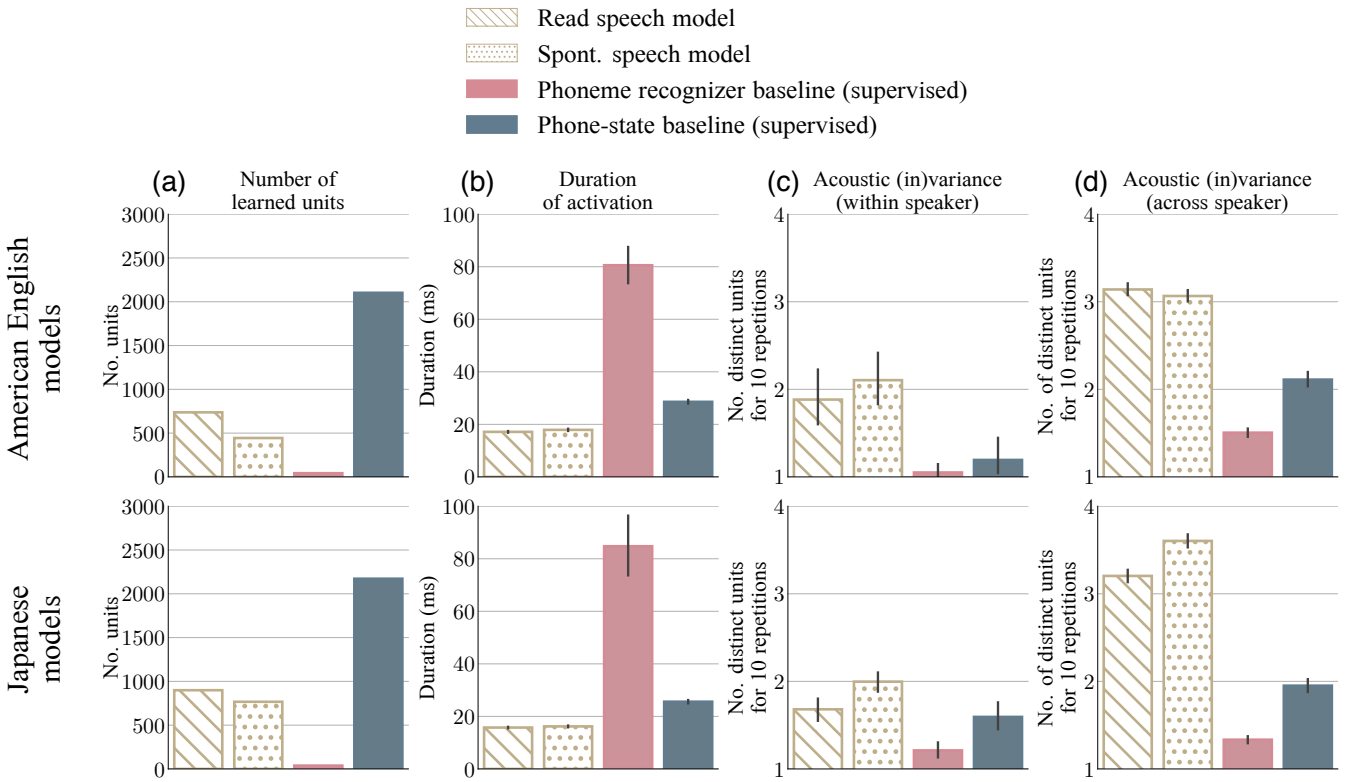

Fig. S7. As in Figure 4, with an additional ASR phone-state baseline (cf. Supplementary Materials and Methods 2). The Gaussian units in the learned (unsupervised) Gaussian mixtures are more similar to the phone-state units than to the phoneme units in the supervised baseline, although some differences remain. Even though the phone states are more numerous than the Gaussian components (a), they remain activated slightly longer on average (b) and they are better aligned with phonetic categories in terms of linguistic content, both within-speakers (c) and across speakers (d).

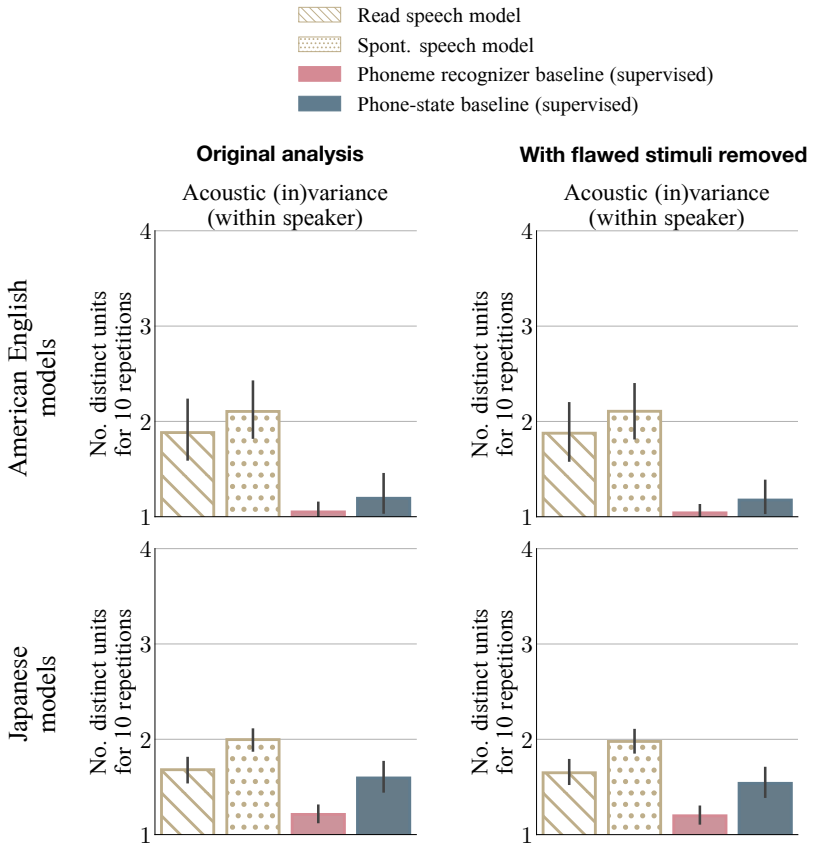

Fig. S8. Supporting evidence for Supplementary Materials and Methods 3. On the left hand side ('Original analysis' panel): acoustic (in)variance analysis for within speaker stimuli as in Figure S7. On the right hand side ('With flawed stimuli removed' panel): same analysis with potentially mispronounced, noisy or misaligned stimuli (as identified through a listening test, see Supplementary Materials and Methods 3) removed. Differences are barely visible and the overall pattern of results remains unchanged.

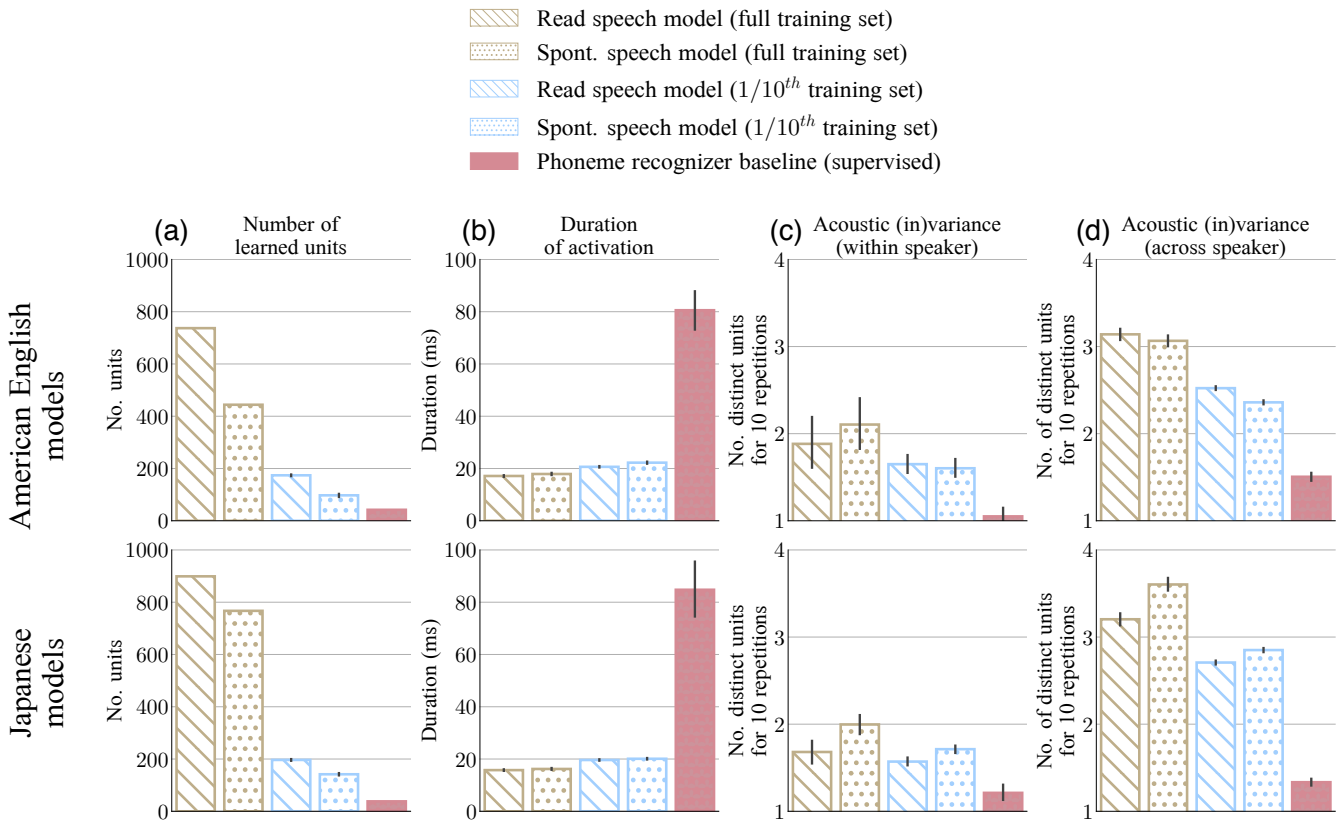

Fig. S9. As in Figure 4, with results for models trained on $1/10^{th}$ subsets of the full training sets added in baby blue (these models already show a reliable cross-linguistic difference in [ɹ]-[l] discriminability between 'American English' and 'Japanese' models, see Figure 3(b)). For the duration and acoustic (in)variance analyses (panels b, c, d), results are averaged over the ten such models trained for each training corpus before standard deviations are estimated. For the number of learned units analysis (panel a), error bars show the standard deviations across the ten trained models. Models trained on $1/10^{th}$ subsets learn much fewer categories (about one fourth as many). This is closer to the typical number of phonemes or of phonetic categories one would expect in a language. Yet, these learned units remain qualitatively different from phonetic categories as shown by the duration and acoustic (in)variance analyses (panels b, c, d). Although their average duration of activation are a few millisecond longer than for models trained on the full training sets, this is still about one fourth of the average duration of speech segments corresponding to phonetic category units. The units learned by the models trained on $1/10^{th}$ subsets also appear slightly more acoustically invariant, with number of distinct units in the acoustic (in)variance tests about 80% that of the models trained on the full training sets (panels c, d). This remains much more variable than the phoneme recognizer baseline, however. Furthermore, for the acoustic (in)variance analysis we have applied a very generous correction for possible misalignment (see Supplementary Materials and Methods 3). This likely causes an overestimation of the acoustic invariance for all the unsupervised models, as indicated by the results on Figure S10. Overall these analyses suggest that the failure of our models to learn phonetic categories cannot be attributed solely to their learning of too many categories.

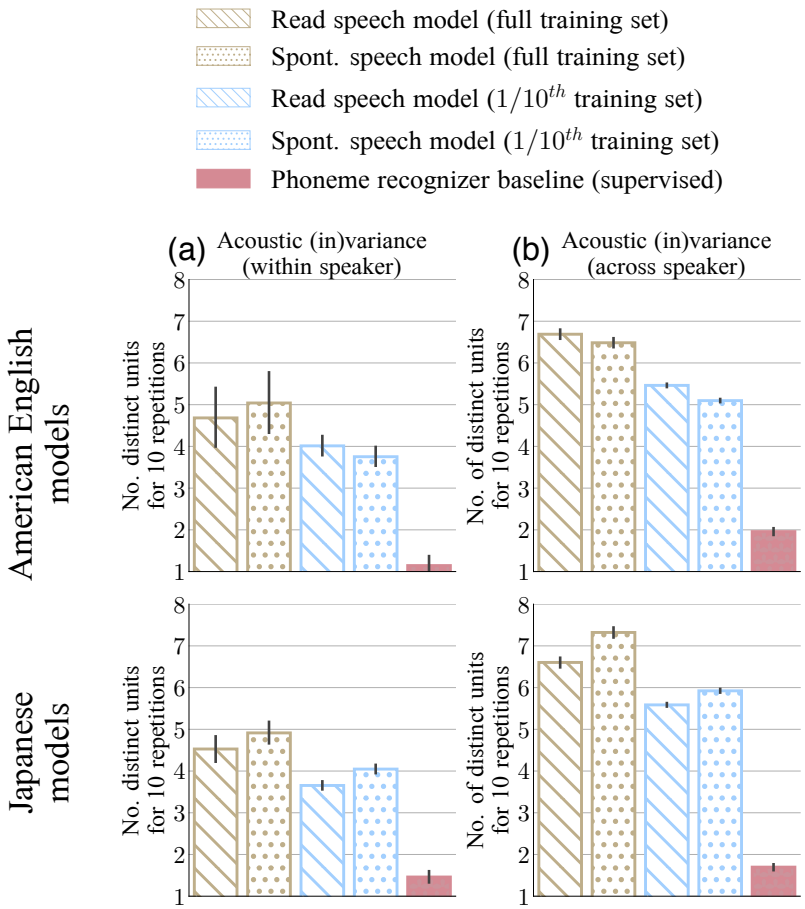

Fig. S10. As in Figure S9 (c, d), but without applying a correction for possible misalignments of the forced-aligned phone centers (Supplementary Materials and Methods 3). For the phoneme recognizer baseline, we see that the average number of distinct units for ten repetitions of a same word shows a small increase compared to the condition with correction for misalignment, with up to about 33% more distinct units (which remains less than what was found for the unsupervised models, *with correction*). In contrast the average number of distinct units more than doubles for our unsupervised models in all cases. This indicates that misalignment of the phone centers is not a very common issue—as the phoneme recognizer baseline manages to find largely invariant units without any correction—suggesting that our main acoustic (in)variance analyses overestimate the acoustic invariance of the units learned by our unsupervised models by a sizable margin.

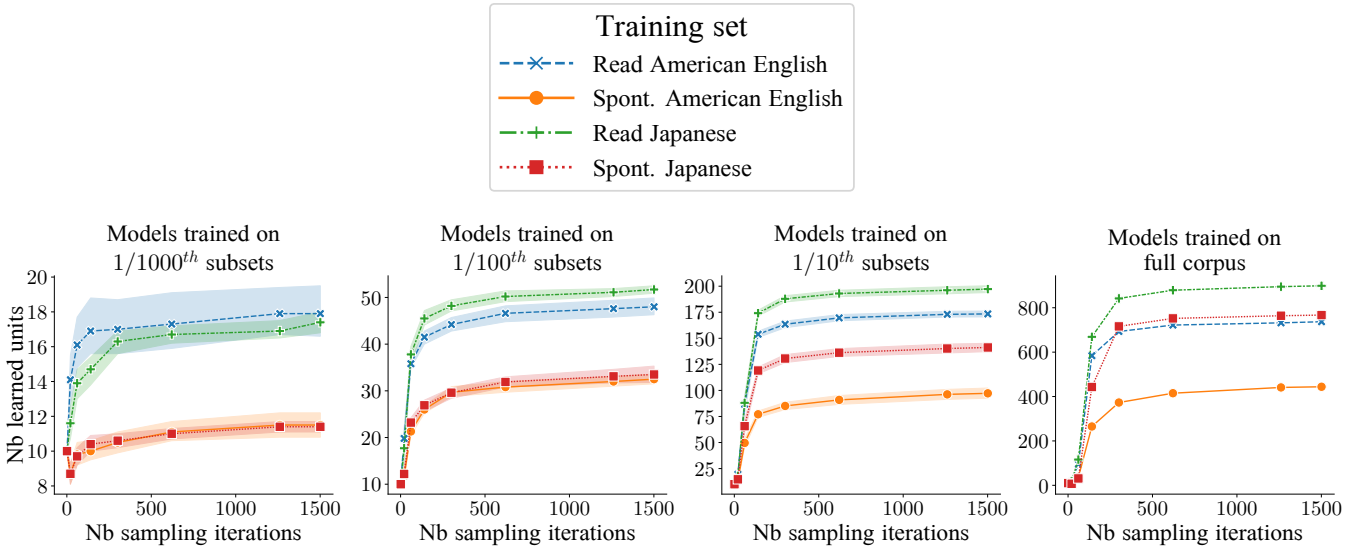

Fig. S11. As a convergence check, we plot the number of learned units (i.e. Gaussian components in the sampled mixture) as a function of the number of sampling iterations. Confidence bands indicate mean +/- one standard deviation in number of learned units for models trained on independent subsets. For models trained on the full corpus no confidence band is available. The number of learned units remains stable after about 600 iterations for all models we trained, suggesting 1500 iterations was enough for our models to converge. For models trained on subsets of the full training set, we also see through the confidence bands that the number of learned categories does not depend a lot on the particular subset selected. Finally, we see evidence that for models trained on small amounts of data, the size of the training set appears to predict the number of learned units well, while for models trained on larger amounts of data, the precise nature of the training set appears to have a stronger effect. Models trained on similar amounts of input (full training sets are about 20 hours long for models trained on read speech and about 10 hours long for model trained on spontaneous speech) learn similar number of categories initially (for $1/1000^{th}$ and $1/100^{th}$ training subsets), but as the size of the training sets gets larger (starting with $1/10^{th}$ training subsets), models trained on Japanese result in larger number of learned categories than models trained on similar amount of American English. This suggests that the number of learned units for the models trained on larger amounts—the models showing cross-linguistic differences in discrimination—does not simply reflect the amount of training input, but also the qualitative characteristics of the training sets.

References

1. D Povey, et al., The kaldi speech recognition toolkit in *Proc. ASRU*. (2011).
2. J Flum, M Grohe, *Parameterized Complexity Theory*. (Springer), pp. 10–17 (2006).
3. Y Benjamini, D Yekutieli, The control of the false discovery rate in multiple testing under dependency. *The annals statistics* **29**, 1165–1188 (2001).
4. J Lee, *U-statistics: Theory and Practice*. (CRC Press), (1990).
5. B De Boer, PK Kuhl, Investigating the role of infant-directed speech with a computer model. *Acoust. Res. Lett. Online* **4**, 129–134 (2003).
6. MH Coen, Self-supervised acquisition of vowels in american english in *Proc. AAAI*. (2006).
7. GK Vallabha, JL McClelland, F Pons, JF Werker, S Amano, Unsupervised learning of vowel categories from infant-directed speech. *Proc. Natl. Acad. Sci.* **104**, 13273–13278 (2007).
8. B McMurray, RN Aslin, JC Toscano, Statistical learning of phonetic categories: insights from a computational approach. *Dev. science* **12**, 369–378 (2009).
9. C Jones, F Meakins, S Muawiyath, Learning vowel categories from maternal speech in gurindji kriol. *Lang. Learn.* **62**, 1052–1078 (2012).
10. F Adriaans, D Swingley, Distributional learning of vowel categories is supported by prosody in infant-directed speech in *Proc. COGSCI*. (2012).
11. B Dillon, E Dunbar, W Idsardi, A single-stage approach to learning phonological categories: Insights from inuktitut. *Cogn. Sci.* **37**, 344–377 (2013).
12. NH Feldman, TL Griffiths, S Goldwater, JL Morgan, A role for the developing lexicon in phonetic category acquisition. *Psychol. review* **120**, 751 (2013).
13. H Rasilo, O Räsänen, UK Laine, Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Commun.* **55**, 909–931 (2013).
14. S Frank, N Feldman, S Goldwater, Weak semantic context helps phonetic learning in a model of infant language acquisition in *Proc. ACL*. (2014).
15. F Adriaans, D Swingley, Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *The J. Acoust. Soc. Am.* **141**, 3070–3078 (2017).
16. F Adriaans, Effects of consonantal context on the learnability of vowel categories from infant-directed speech. *The J. Acoust. Soc. Am.* **144**, EL20–EL25 (2018).
17. RAH Bion, K Miyazawa, H Kikuchi, R Mazuka, Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS ONE* **8**, e51594 (2013).
18. S Antetomaso, et al., *Modeling phonetic category learning from natural acoustic data*. (Cascadilla Press), (2017).
19. K Miyazawa, H Kikuchi, R Mazuka, Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model in *Proc. INTERSPEECH*. (2010).
20. K Miyazawa, H Miura, H Kikuchi, R Mazuka, The multi timescale phoneme acquisition model of the self-organizing based on the dynamic features in *Proc. INTERSPEECH*. (2011).
21. FH Guenther, MN Gjaja, The perceptual magnet effect as an emergent property of neural map formation. *The J. Acoust. Soc. Am.* **100**, 1111–1121 (1996).
22. H Chen, CC Leung, L Xie, B Ma, H Li, Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study in *Proc. ISCA*. (2015).
23. J Chang, JW Fisher III, Parallel sampling of dp mixture models using sub-cluster splits in *Proc. NEURIPS*. (2013).
24. PK Kuhl, et al., Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philos. Transactions Royal Soc. B: Biol. Sci.* **363**, 979–1000 (2007).
25. G Dehaene-Lambertz, The human infant brain: A neural architecture able to learn language. *Psychon. bulletin & review* **24**, 48–55 (2017).
26. E Hermann, S Goldwater, Multilingual bottleneck features for subword modeling in zero-resource languages in *Proc. INTERSPEECH*. (2018).
27. K Chládková, N Paillereau, The what and when of universal perception: A review of early speech sound acquisition. *Lang. Learn.* **n/a** (2020).
28. K Behnke, *The Acquisition of Phonetic Categories in Young Infants: A Self-Organizing Artificial Neural Network Approach*, MPI series in psycholinguistics. (MPI, Nijmegen), (1998).
29. PK Kuhl, et al., Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. science* **9**, F13–F21 (2006).
30. T Tsushima, et al., Discrimination of english/rl/and/wy/by japanese infants at 6-12 months: language-specific developmental changes in speech perception abilities in *Proc. ICSLP*. (1994).
31. K Idemaru, LL Holt, The developmental trajectory of children's perception and production of english/r/-/l. *The J. Acoust. Soc. Am.* **133**, 4232–4246 (2013).
32. S Tsuji, A Cristia, Perceptual attunement in vowels: A meta-analysis. *Dev. psychobiology* **56**, 179–191 (2014).
33. A Gagliardi, J Lidz, Statistical insensitivity in the acquisition of Tsez noun classes. *Language* **90**, 58–89 (2014).
34. R Mugitani, et al., Perception of vowel length by japanese-and english-learning infants. *Dev. psychology* **45**, 236 (2009).
35. Y Sato, Y Sogabe, R Mazuka, Discrimination of phonemic vowel length by japanese infants. *Dev. Psychol.* **46**, 106 (2010).

36. DK Burnham, Developmental loss of speech perception: Exposure to and experience with a first language. *Appl. Psycholinguist.* **7**, 207–240 (1986).
37. V Hazan, S Barrett, The development of phonemic categorization in children aged 6–12. *J. phonetics* **28**, 377–396 (2000).
38. R Li, T Schatz, Y Matusevych, S Goldwater, NH Feldman, Input matters in the modeling of early phonetic learning in *Proc. COGSCI.* (2020).
39. E Dunbar, et al., The zero resource speech challenge 2019: TTS without T. *CoRR* **abs/1904.11469** (2019).
40. DB Pisoni, Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Percept. Psychophys.* **13**, 253–260 (1973).
41. WT Fitch, *The evolution of language.* (Cambridge University Press), (2010).
42. LS Scott, O Pascalis, CA Nelson, A domain-general theory of the development of perceptual discrimination. *Curr. directions psychological science* **16**, 197–201 (2007).
43. D Maurer, JF Werker, Perceptual narrowing during infancy: A comparison of language and faces. *Dev. Psychobiol.* **56**, 154–178 (2014).
44. DJ Kelly, et al., The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychol. Sci.* **18**, 1084–1089 (2007).
45. RH Friendly, D Rendall, LJ Trainor, Learning to differentiate individuals by their voices: Infants' individuation of native-and foreign-species voices. *Dev. psychobiology* **56**, 228–237 (2014).
46. DJ Levitin, RJ Zatorre, On the nature of early music training and absolute pitch: A reply to brown, sachs, cammuso, and folstein. *Music. Perception: An Interdiscip. J.* **21**, 105–110 (2003).
47. FA Russo, DL Windell, LL Cuddy, Learning the ""special note"": Evidence for a critical period for absolute pitch acquisition. *Music. Perception: An Interdiscip. J.* **21**, 119–127 (2003).
48. EE Hannon, SE Trehub, Tuning in to musical rhythms: Infants learn more readily than adults. *Proc. Natl. Acad. Sci.* **102**, 12639–12643 (2005).
49. SB Palmer, L Fais, RM Golinkoff, JF Werker, Perceptual narrowing of linguistic sign occurs in the 1st year of life. *Child development* **83**, 543–553 (2012).
50. S Bao, Perceptual learning in the developing auditory cortex. *Eur. J. Neurosci.* **41**, 718–724 (2015).
51. EJ Yang, EW Lin, TK Hensch, Critical period for acoustic preference in mice. *Proc. Natl. Acad. Sci.* **109**, 17213–17220 (2012).
52. EA Simpson, et al., Face detection and the development of own-species bias in infant macaques. *Child development* **88**, 103–113 (2017).
53. WM Weikum, TF Oberlander, TK Hensch, JF Werker, Prenatal exposure to antidepressants and depressed maternal mood alter trajectory of infant speech perception. *Proc. Natl. Acad. Sci.* **109**, 17221–17227 (2012).
54. J Gervain, et al., Valproate reopens critical-period learning of absolute pitch. *Front. systems neuroscience* **7**, 102 (2013).
55. JF Werker, TK Hensch, Critical periods in speech perception: new directions. *Annu. review psychology* **66**, 173–196 (2015).
56. M Versteegh, X Anguera, A Jansen, E Dupoux, The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Comput. Sci.* **81**, 67–72 (2016).
57. M Heck, S Sakti, S Nakamura, Unsupervised linear discriminant analysis for supporting dpgmm clustering in the zero resource scenario. *Procedia Comput. Sci.* **81**, 73–79 (2016).
58. M Heck, S Sakti, S Nakamura, Feature optimized dpgmm clustering for unsupervised subword modeling: A contribution to zerospeech 2017 in *Proc. ASRU.* (2017).
59. PW Jusczyk, *Developing Phonological Categories from the Speech Signal.* (York, Timonium, MD), pp. 17–64 (1992).
60. PW Jusczyk, From general to language-specific capacities: the WRAPSA model of how speech perception develops. *J. Phonetics* **21**, 3–28 (1993).
61. P Jusczyk, The discovery of spoken language (1997).
62. E Dupoux, Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition* **173**, 43–59 (2018).
63. ES Spelke, SA Lee, Core systems of geometry in animal minds. *Philos. Transactions Royal Soc. B: Biol. Sci.* **367**, 2784–2793 (2012).
64. DL Yamins, JJ DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. neuroscience* **19**, 356–365 (2016).
65. C Zhuang, et al., Unsupervised neural network models of the ventral visual stream. bioRxiv (2020).
66. KL Stachenfeld, MM Botvinick, SJ Gershman, The hippocampus as a predictive map. *Nat. neuroscience* **20**, 1643 (2017).
67. JC Whittington, et al., The tolman-eichenbaum machine: Unifying space and relational memory through generalisation in the hippocampal formation. bioRxiv (2019).
68. C Frith, *Making up the mind: How the brain creates our mental world.* (John Wiley & Sons), (2013).
69. B Gauthier, R Shi, Y Xu, Learning phonetic categories by tracking movements. *Cognition* **103**, 80–106 (2007).
70. ES Levy, W Strange, Effects of consonantal context on perception of french rounded vowels by american english adults with and without french language experience. *The J. Acoust. Soc. Am.* **111**, 2361–2362 (2002).
71. NA Macmillan, CD Creelman, *Detection theory: A user's guide.* (Psychology press), (2004).
72. NH Feldman, TL Griffiths, JL Morgan, The influence of categories on perception: Explaining the perceptual magnet effect

as optimal statistical inference. *Psychol. review* **116**, 752 (2009).
73. JF Werker, S Curtin, Primir: A developmental framework of infant speech processing. *Lang. learning development* **1**, 197–234 (2005).
74. T Schatz, Ph.D. thesis (Université Paris 6) (2016).
75. H Hofmann, K Kafadar, H Wickham, Letter-value plots: Boxplots for large data, (had.co.nz), Technical report (2011).